AI 3000 (CS 5500) : Reinforcement Learning

Easwar Subramanian

TCS Innovation Labs, Hyderabad

Email : easwar.subramanian@tcs.com / cs5500.2020@iith.ac.in

August 19, 2021

Overview

1. Introduction
2. RL : Framework, Components and Challenges
3. Historical Notes
4. Motivation and Success Stories
5. Course Logistics

Introduction

Machine Learning

" Machine learning is about developing bots that has the ability to automatically learn and improve from experience without being explicitly programmed "

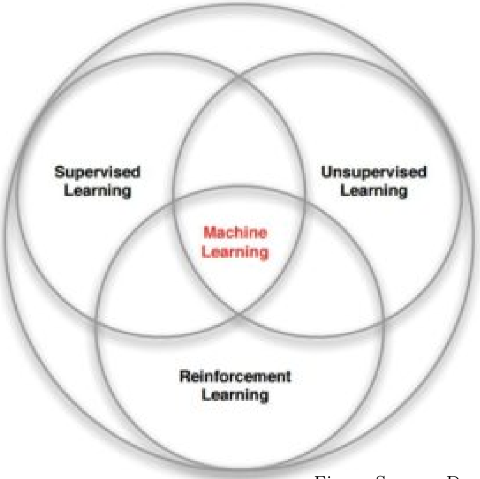
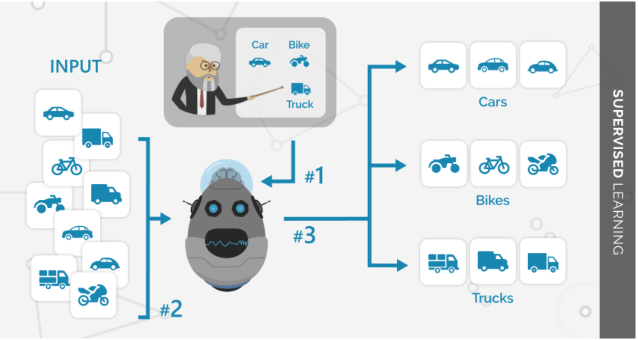


Figure Source: David Silver's RL course

Supervised Learning

- **Data** : $(x, y) \rightarrow x$ is data and y is label
- **Goal**: Learn a function f to map $y = f(x)$
- **Problems** : Classification or Regression

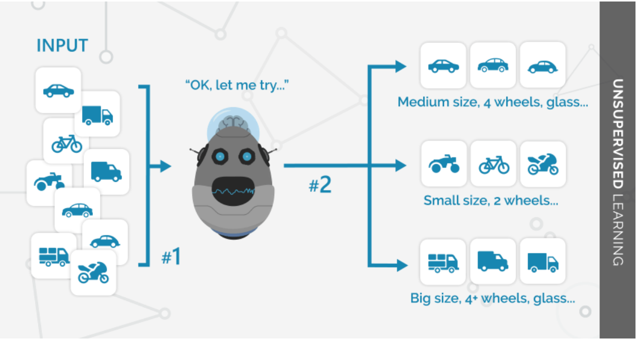


Classification

Figure Source: Aura Portal - AI/ML Blog

Unsupervised Learning

- **Data** : $(x) \rightarrow$ Only data; No label
- **Goal**: Learn underlying structure
- **Techniques** : Clustering

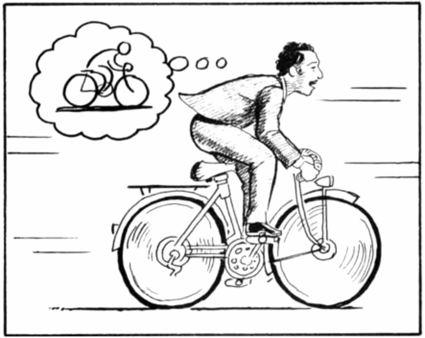


Clustering

Figure Source: Aura Portal - AI/ML Blog

Reinforcement Learning

- **Data** : Agent interacts with environment to collect data
- **Goal** : Agent learns to interact with environment to maximize an utility
- **Examples** : Learn a task, Navigation

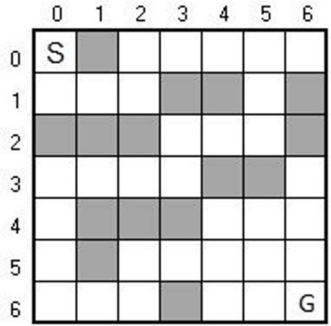


Learn to cycle (task)

Figure Source: worldmodels.github.io

Example : Navigation

- **Task :** Start from square S and reach square G in as less moves as possible



Navigation in grid world

- One has to make **sequence** of moves (actions)
- Action chosen **determine** which squares (states) would be visited subsequently
- Reaching the **goal state** will fetch a reward; Visiting Intermediate squares (states) may or may not fetch reward

Figure Source: Genevieve Hayes : Medium Post

Sequential Decision Making

Supervised or Unsupervised Setting

- System is making a isolated decision; i.e., classification, regression or clustering;
- Decision does not affect future observations

Reinforcement Learning

- Generally, the agent makes a sequence of decisions (or actions)
- Actions affect future observations
- Actions taken have consequences

Types of Learning : Summary

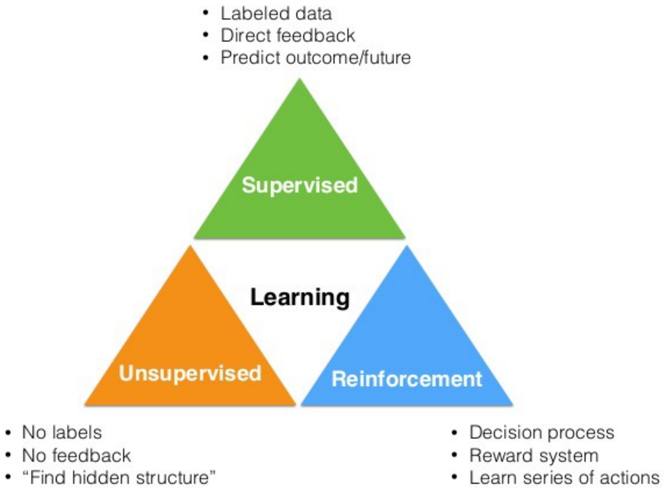
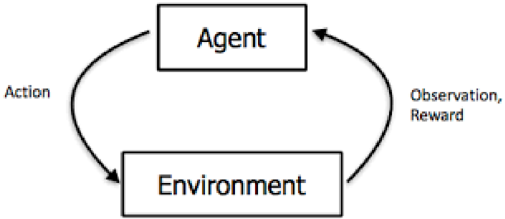


Figure Source: Saggie

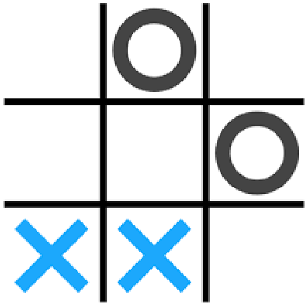
RL : Framework, Components and Challenges

Reinforcement Learning : Framework



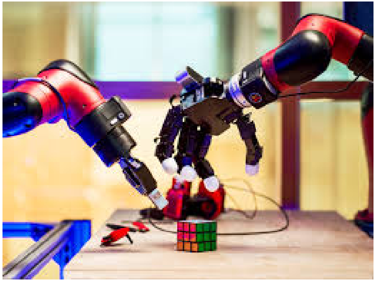
- Observations are non i.i.d and are sequential in nature
- Agent's action (may) affect the subsequent observations seen
- There is no supervisor; Only reward signal (feedback)
- Reward or feedback can be delayed

Example : Tic-Tac-Toe



- Observations : Board position
- Actions : Moves
- Reward : Win or Loss

Example : Robotics



- Observations : Image from in-built camera
- Actions : Motor current for movement
- Reward : Task success measure

Example : Inventory Control

- **Observations** : Stock levels
- **Actions** : What to purchase
- **Reward** : Profit

Components of RL : Agent and Environment

Agent

- A system that takes actions to change the state of the environment (Decision maker)
- Executes action upon receiving observation
- For taking an action the agent receives an appropriate reward

Environment

- An **external system** that an agent can perceive and act on.
- Receives action from agent and in response emits appropriate reward and (next) observation

Components of RL : State and Reward

State

- State can be viewed as a summary or an abstraction of the past history of the system
  - ★ For example, in Tic-Tac-Toe, the state could be raw image or vector representation of the board

Reward

- Reward is a scalar feedback signal
- Indicates how well agent acted at a certain time
- The agent's aim is to maximise cumulative reward

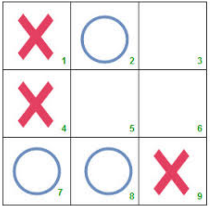
Slide Credit: David Silver's RL Course

Reinforcement Learning : Challenges

- Delayed Feedback
- Credit Assignment Problem
- Stochastic Environment
- Definition of Reward Function
- Data Collection Problem

Historical Notes

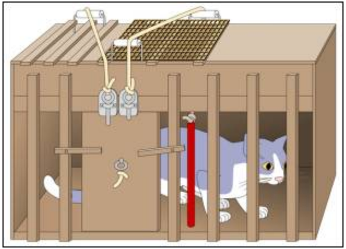
Learning by Trial and Error



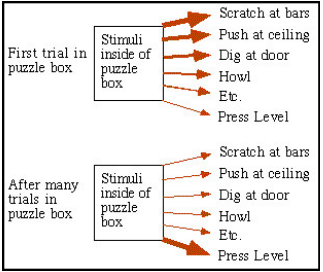
Tic-Tac-Toe

- Random movements by agent is akin to exploration
- Exploration can help the agent place 'X' in square number 5
- Reward obtained from placing 'X' in square number 5 can now be remembered in terms of updating the policy or value function

Thondrike's Cat : Psychophysical Experiment



Thondrike's cat



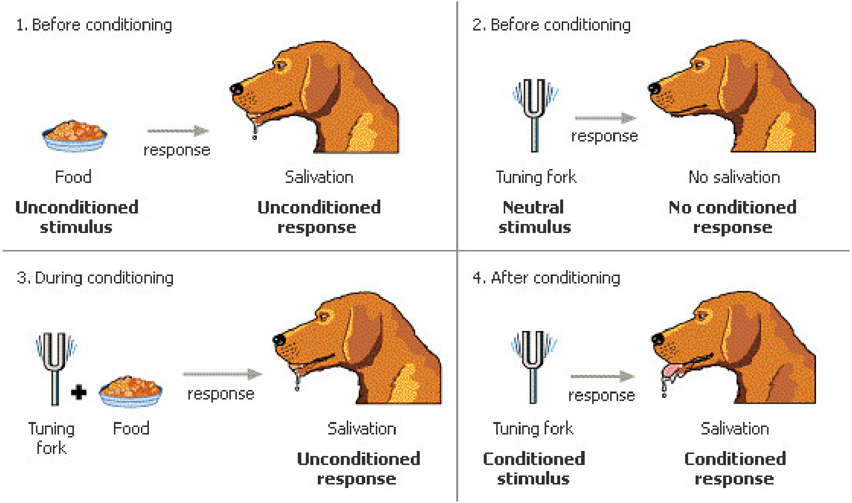
Law of Effect

Law of Effect (1898)

Any behaviour that is followed by pleasant consequences is likely to be repeated, and any behaviour followed by unpleasant consequences is likely to be stopped

Figure Source: Oscar Education : Blogpost

Pavlov's Dog



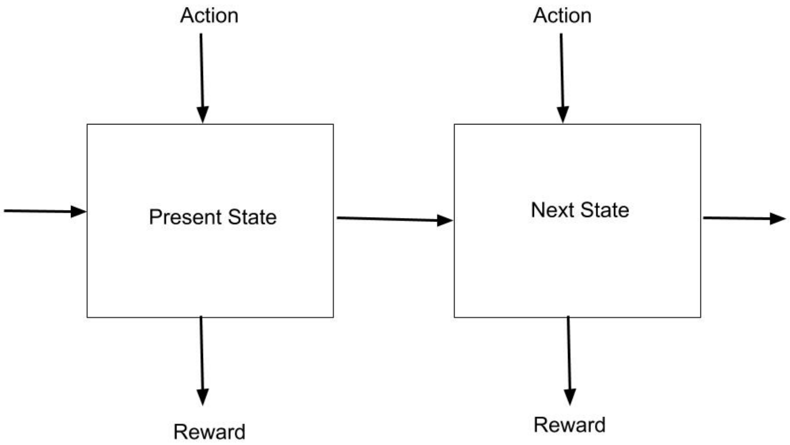
Pavlov's Dog

Figure Source: https://www.age-of-the-sage.org/psychology/pavlov.html

Connections to Temporal Difference

- Ivav Pavlov laid the ground for classical conditioning (1901)
- First theory that incorporated time into the learning procedure
- **Rescorla-Wagner** (RW) (1972) model is a formal model to explain Pavlovian conditioning
- Temporal-Difference (TD) learning, that extends RW model, is an approach to learning how to predict a quantity that depends on future values of a given signal (Sutton, 1984)
- TD learning forms the basis of almost all RL algorithms that we see today

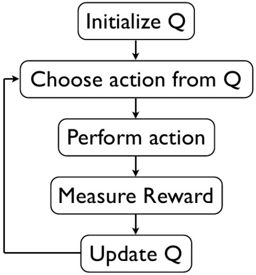
Connections to Optimal Control



Connections to Optimal Control

- Outcomes are partly random and partly under the control of the decision maker
- Markov Decision Process (MDP) (Bellman, 1957) is used as a framework to model and solve sequential decision problem
- People working in control theory have contributed to optimal sequential decision making

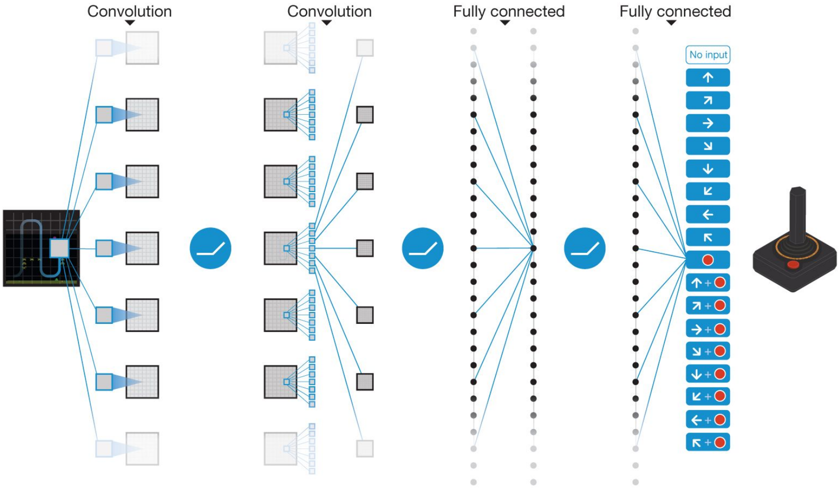
Modern Reinforcement Learning



- The temporal difference (TD) thread and the optimal control thread were bought together by Watkins (1989) when he proposed the famous **Q-learning algorithm**
- Gerald Tesauro (1992) employed TD learning to play **backgammon**; The developed software agent was able to beat experts

Figure Source: https://www.linuxjournal.com/article/11038

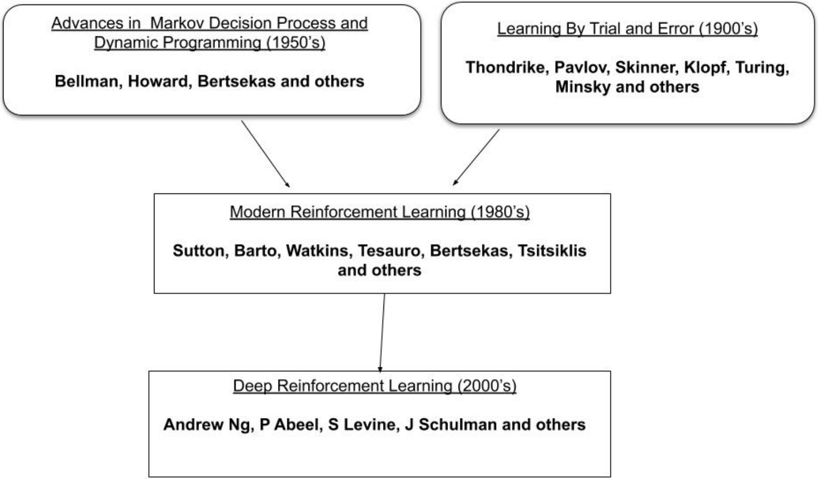
Era of Deep (Reinforcement) Learning



Deep Neural Net for Atari Games

Figure Source: Minh et.al, 2015

Reinforcement Learning : History



Motivation and Success Stories

Motivation

Why study Reinforcement Learning (RL) now ?

- Advances in computational capability
- Advances in deep learning
- Advances in reinforcement learning
  - ★ Subject matter of this course !

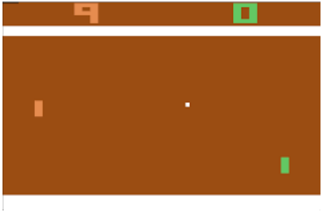
Slide Credit: Sergey Levine course on Deep RL at UCB

Sucess Stories

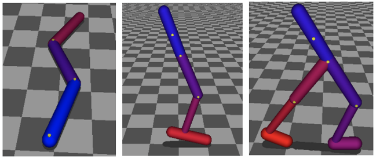
(a) Ng et al 2004

(b) Kohl et al 2004

Sucess Stories



(c) Minh et al 2013



(d) Schulman et al 2016

(d) Silver et al. 2016

Towards Intelligent Systems

- Things that we can all do (Walking) (Evolution, may be)
- Things that we learn (driving a bicycle, car etc)
- We learn a huge variety of things (music, sport, arts etc)
- We can learn 'difficult' tasks as well

We are still far from building a 'reasonable' intelligent system

- We are taking baby steps towards the goal of building intelligent systems
- **Reinforcement Learning (RL) is one of the important paradigm towards that goal**

Course Logistics

Course Content - Part A

Modern Reinforcement Learning

- Markov Decision Process
- Dynamic Programming and Bellman Optimality Principle
- Value and Policy Iteration
- Convergence Properties of Value and Policy Iteration
- Model Free Prediction
- Model Free Control : Q-Learning and SARSA

Course Content - Part B

Deep Reinforcement Learning

- Deep Q-Learning and Variants
- Policy Gradient Approaches
- Variance Reduction in Policy Gradient Methods
- Actor Crtic Algorithms
- Deterministic Policy Gradients
- Advanced Policy Gradient Methods : TRPO and PPO

Venue and Timing

- **Mode**
  - ★ GMeet
- **Timing - Slot P**
  - ★ Monday - 2.30 PM to 4.00 PM
  - ★ Thursday - 4.00 PM to 5.30 PM

Course Material : Books

- Reinforcement Learning : Sutton and Barto
- Reinforcement Learning and Optimal Control, Bertsekas and Tsitsiklis
- Dynamic Programming and Optimal Control (I and II) by Bertsekas

Course Material : Online Material

- David Silver's course on Reinforcement Learning
- Stanford course on Deep RL (Sergey Levine)
- Deep RL BootCamp (Pieter Abeel)
- John Schulman's lectures on Policy Gradient Methods
- ... and many others

Course Material : From India

- Prof. B. Ravindran's Course on RL (NPTEL)
- Dr. Abir Das's Course on RL (IIT KGP)

Course Prerequisites

- **Necessary Prerequisites**
  - ★ Probability
  - ★ Linear Algebra
  - ★ Machine Learning
- **Desirable Prerequisites**
  - ★ Deep Learning
- **Programming Prerequisites**
  - ★ Good Proficiency in Python
  - ★ Tensorflow / Theano / PyTorch / Keras
  - ★ Other Associated Python Libraries

Course Evaluation - Tentative

- **Homeworks / Assignments** : Four or FIve in Total (30 - 40 %)
- **Exams / Quiz** : Two or Three in Total (60 - 70 %)
- **Course Project** : Details will follow (In lieu of some assignments / exam)

Attribution and Disclaimer

- Most concepts, ideas and figures, that form part of course lectures, are from several sources from across web; Most of them are listed as course material
- Care is taken to provide appropriate attribution; Omissions, if any, are regretted and unintentional
- Material prepared only for <u>learning / teaching</u> purpose
- Original <u>authorship / copyright</u> rests with the respective authors / publishers

Enjoy Learning !